

AMENDMENTS TO THE CLAIMS

This listing of claims will replace all prior versions of claims in the application:

Listing of Claims:

1. (Currently Amended) A system that facilitates incremental web crawls comprising the following components stored in computer memory and executable by a processor:
an indexer that places items with similar properties into respective chunks, wherein the items are the results returned by a web crawl; and,
a chunk map that stores at least some of the properties associated with the respective chunk, the stored properties are shared by all the items in the respective chunk, wherein the properties are at least one of average time between change ~~or~~ and average importance of documents ~~comprising a particular~~ in the respective chunk, the chunk map employed to facilitate an incremental web re-crawl, wherein the properties of each chunk stored in the chunk map are utilized to determine a re-crawl of all the items in that chunk.
2. (Original) The system of claim 1, the items comprising information associated with a Uniform Resource Locator.
3. (Original) The system of claim 1, the items comprising at least one of an HTML file, a PDF file, a PS file, a PPT file, an XLS file and a DOC file.
4. (Original) The system of claim 1, the items receives from a crawler, the crawler responsible for a specific set of Uniform Resource Locators.
5. (Original) The system of claim 1, further comprising a master control process that can modify the chunk map to facilitate load balancing amongst a plurality of crawlers.
6. (Original) The system of claim 1, further comprising a master control process that serves as an interface between a crawler and a re-crawl controller.

7. (Original) The system of claim 6, wherein the master control process maintains a known chunks table that stores information for components of a system.
8. (Original) The system of claim 6, wherein the master control process exposes an interface for communication with a component of the system.
9. (Original) The system of claim 8, wherein the interface returns a list of chunks the component should have and where to get the chunks.
10. (Original) The system of claim 8, wherein the interface returns a list of the chunks that should be actively served by the component.
11. (Original) The system of claim 8, wherein the interface returns a range of chunk identifiers to use in building a new chunk by the component.
12. (Original) The system of claim 8, wherein the interface causes an old chunk to be retired by the system.
13. (Original) The system of claim 6, wherein the master control process facilitates movement of chunks from one component to another component.
14. (Original) The system of claim 13, wherein movement of chunks is based, at least in part, upon at least one of rebalancing index servers after one goes down, re-crawling pages previously crawled, and, restoring a state of a crawler after it has crashed.
15. (Original) The system of claim 1, further comprising a re-crawl component that employs the chunk map to determine which chunks, if any, to re-crawl at a particular time.
16. (Cancelled)

17. (Original) The system of claim 1, further comprising an index chunk that stores information associated with an index of at least some of the items.
18. (Original) The system of claim 1, further comprising a rank chunk that stores a static rank associated with an index chunk.
19. (Currently Amended) A method of performing document re-crawl comprising:
parsing a first chunk for uniform resource locators, wherein a chunk map that stores properties associated with the respective chunk is employed to determine the first chunk, wherein the stored properties are shared by all the items in the respective chunk[[,]];
re-crawling the uniform resource locators; and,
forming a second chunk separate from the first chunk, based at least in part, upon the re-crawled uniform resource locators.
20. (Original) The method of claim 19 comprising at least one of the following acts:
determining whether any chunks are to be retired;
moving the first chunk; and,
destroying the first chunk.
21. (Original) One or more computer readable media having stored thereon computer executable instructions for carrying out the method of claim 19.
22. (Currently Amended) A method of performing document re-crawl comprising:
parsing a first chunk for uniform resource locators, wherein the uniform resource locators are stored as a result of one or more web crawls;
accessing a chunk map ~~containing~~ comprising properties associated with the respective chunk[[s]] ~~of data as a result of one or more web crawls~~, the stored properties are shared by all the items in the respective chunk, wherein the properties are at least one of average time between change or average importance of documents ~~comprising a particular~~ in the respective chunk; and,
periodically determining, based on the properties of each chunk in the chunk map, whether to re-crawl the chunk of data.

23. (Original) The method of claim 22, the period determination being based, at least in part, upon, at least one of average time between change and average importance of documents comprising a particular chunk.

24. (Currently Amended) A computer-readable medium encoded with a data ~~packet~~ structure transmitted between two or more computer components that facilitates document re-crawl, the data ~~packet~~ structure comprising:

a chunk header that includes metadata associated with the data structure ~~packet~~, the metadata shared by all the items in ~~the~~ a chunk that comprises document files associated with one or more uniform resource locators;

an offset section that provides offset information associated with the document files; and,
the document files that include content found on the Internet, wherein the average of the at least one of the properties of all the document files determines if the document should be re-crawled.

25. (Currently Amended) The data structure ~~packet~~ of claim 24, at least one of the document files comprising at least one of an HTML file, a PDF file, a PS file, a PPT file, an XLS file and a DOC file.

26. (Currently Amended) A system that facilitates increment web crawls comprising the following components stored in computer memory and executable by a processor:

means for placing items with similar properties into respective chunks, wherein the items are results returned by a web crawl; and,

means for storing at least some of the properties associated with the respective chunk, wherein the stored properties are shared by all the items in the respective chunk, and the properties are at least one of average time between change or average importance of documents ~~comprising a particular~~ in the respective chunk, and employing the stored properties of each chunk to facilitate an incremental web re-crawl.

27. (Original) The system of claim 26, the items comprising information associated with a Uniform Resource Locator.

28. (Original) The system of claim 26, the items comprising at least one of an HTML file, a PDF file, a PS file, a PPT file, an XLS file and a DOC file.